Title

YeATS - a tool suite for analyzing RNA-seq derived transcriptome identifies a highly transcribed putative extensin in heartwood/sapwood transition zone in black walnut.

Authors

Chakraborty, Sandeep
Britton, Monica
Wegrzyn, Jill
et al.

Publication Date

2015

DOI

10.12688/f1000research.6617.2

Peer reviewed

RESEARCH ARTICLE

REVISED YeATS - a tool suite for analyzing RNA-seq derived transcriptome identifies a highly transcribed putative extensin in heartwood/sapwood transition zone in black walnut [version 2; referees: 3 approved]

Sandeep Chakraborty[1], Monica Britton[2], Jill Wegrzyn[3], Timothy Butterfield[1], Pedro José Martínez-García[1], Russell L. Reagan[1], Basuthkar J. Rao[4], Charles A. Leslie[1], Mallikarjuna Aradhaya[1], David Neale[1], Keith Woeste[5], Abhaya M. Dandekar[1]

[1]Plant Sciences Department, University of California, Davis, CA, 95616, USA
[2]UC Davis Genome Center Bioinformatics Core Facility, University of California, Davis, CA, 95616, USA
[3]Department of Ecology and Evolutionary Biology, University of Connecticut, Storrs, CT, 06269, USA
[4]Department of Biological Sciences, Tata Institute of Fundamental Research, Homi Bhaba Road, Mumbai, 400, India
[5]USDA Forest Service Hardwood Tree Improvement and Regeneration Center, Purdue University, West Lafayette, IN, 47907, USA

v2 **First published:** 17 Jun 2015, **4**:155 (doi: 10.12688/f1000research.6617.1)
Latest published: 06 Nov 2015, **4**:155 (doi: 10.12688/f1000research.6617.2)

Abstract

The transcriptome provides a functional footprint of the genome by enumerating the molecular components of cells and tissues. The field of transcript discovery has been revolutionized through high-throughput mRNA sequencing (RNA-seq). Here, we present a methodology that replicates and improves existing methodologies, and implements a workflow for error estimation and correction followed by genome annotation and transcript abundance estimation for RNA-seq derived transcriptome sequences (YeATS - Yet Another Tool Suite for analyzing RNA-seq derived transcriptome). A unique feature of YeATS is the upfront determination of the errors in the sequencing or transcript assembly process by analyzing open reading frames of transcripts. YeATS identifies transcripts that have not been merged, result in broken open reading frames or contain long repeats as erroneous transcripts. We present the YeATS workflow using a representative sample of the transcriptome from the tissue at the heartwood/sapwood transition zone in black walnut. A novel feature of the transcriptome that emerged from our analysis was the identification of a highly abundant transcript that had no known homologous genes (GenBank accession: KT023102). The amino acid composition of the longest open reading frame of this gene classifies this as a putative extensin. Also, we corroborated the transcriptional abundance of proline-rich proteins, dehydrins, senescence-associated proteins, and the DNAJ family of chaperone proteins. Thus, YeATS presents a workflow for analyzing RNA-seq data with several innovative features that differentiate it from existing software.

Open Peer Review

Referee Status: ✓ ✓ ✓

	Invited Referees 1	2	3
REVISED version 2 published 06 Nov 2015	✓ report	✓ report	✓ report
version 1 published 17 Jun 2015	? report		

1 **Varodom Charoensawan**, Mahidol University Thailand

2 **Michael Love**, Harvard TH Chan School of Public Health USA

3 **Binay Panda**, Institute of Bioinformatics and Applied Biotechnology(IBAB) India

Discuss this article

Comments (0)

Corresponding author: Sandeep Chakraborty (sanchak@gmail.com)

How to cite this article: Chakraborty S, Britton M, Wegrzyn J *et al.* **YeATS - a tool suite for analyzing RNA-seq derived transcriptome identifies a highly transcribed putative extensin in heartwood/sapwood transition zone in black walnut [version 2; referees: 3 approved]** *F1000Research* 2015, **4**:155 (doi: 10.12688/f1000research.6617.2)

Grant information: The authors wish to acknowledge support from the California Walnut Board and UC Discovery program.
The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: No competing interests were disclosed.

First published: 17 Jun 2015, **4**:155 (doi: 10.12688/f1000research.6617.1)

REVISED **Amendments from Version 1**

In this version, we have

1) Added two new authors based on their inputs to the manuscript

2) Provided IDs to the submissions of the transcriptome(s).

3) Created github repository with README. It is to be noted that this is not meant to be a software article, so the software provided is not release quality. https://github.com/sanchak/YEATSCODE1

4) Incorporated several minor points raised by reviewer.

See referee reports

Introduction

Analysis of the complete set of RNA molecules in a cell, the transcriptome, is critical to understanding the functional aspects of the genome of an organism. Most transcripts get translated into proteins by the ribosome[1]. Non-translated transcripts (noncoding RNAs) may be alternatively spliced and/or broken into smaller RNAs, the importance of which have only recently been recognized[2]. Transcriptional levels vary significantly based on environmental cues[3], and/or disease[4]. Quantifying transcriptional levels constitutes an important methodology in current biological research. Traditional methods like RNA:DNA hybridization[5] and short sequence-based approaches[6] have been supplanted recently by a high-throughput DNA sequencing method - RNA-seq[7,8]. Concomitant with the introduction of RNA-seq has been the development of a diverse set of computational methods for analyzing the resultant data[9–21].

In the current work, we present a methodology for analyzing RNA-seq data that has been assembled into transcripts (**YeATS** - **Ye**t **A**nother **T**ool **S**uite for analyzing RNA-seq derived transcriptome). The process of associating genomic open reading frames (ORF) to a set of transcripts (transcriptome) is the key step in YeATS, enabling identification and correction of specific errors arising from sequencing and/or assembly, a novel feature missing in most known tools. These errors include transcripts that have not been merged, a transcript having broken ORFs and transcripts containing long repeats. Also, YeATS identifies noncoding RNAs by comparison to compiled databases[22], transcripts with multiple coding sequences and highly transcribed genes (based on simple normalization of raw counts followed by sorting).

Here, the YeATS workflow is demonstrated using a representative sample of the transcriptome from the tissue at the heartwood/sapwood transition zone in black walnut (*Juglans nigra* L.). We have identified transcripts that have sequencing and/or assembly errors (~5%). A novel feature that emerged from our analysis was the presence of a highly transcribed gene that had no known homologous counterpart in the entire BLAST database. The amino acid composition of the longest open reading frame of this gene consists of a high percentage of leucine, histidine and valine, and classifies this as a putative extensin[23]. Given the economic and ecological importance of black walnut timber, characterization of such genes will enhance our understanding of the mechanisms underlying the unique properties associated with the wood of these trees[24]. The significance of proline-rich proteins[25], dehydrins[26], senescence-associated proteins[27] and DNAJ[28] proteins to the formation of heartwood was established through their transcriptional abundance. Finally, based on transcripts that have no known homologs, we have identified noncoding RNAs by comparison with the noncoding RNA database for *Arabidopsis*[22]. Thus, in the current work, we present a workflow (YeATS) with several novel features absent in most currently available software.

Methods

In silico methods

The input to YeATS is a set of post assembly transcripts as a fasta file (ϕ_{TRS}). The first step is to identify the set of genes (proteins) encoded by ϕ_{TRS}. This is done by associating a proper open reading frame (ORF) to each transcript. This involves a comprehensive automated BLAST run[29].

For each transcript in ϕ_{TRS}, we generate the three longest ORFs (using the 'getorf' utility in the EMBOSS suite[30]) (Figure 1). These three ORFs are BLAST'ed to the full non-redundant protein sequences ('nr') database. For a given E-value cutoff (1E-12 in the current work), we create four sets

1. Only one ORF is less than the cutoff - the transcript is uniquely annotated.
2. None of the ORFs is less than the cutoff - the transcript has no known homologs.

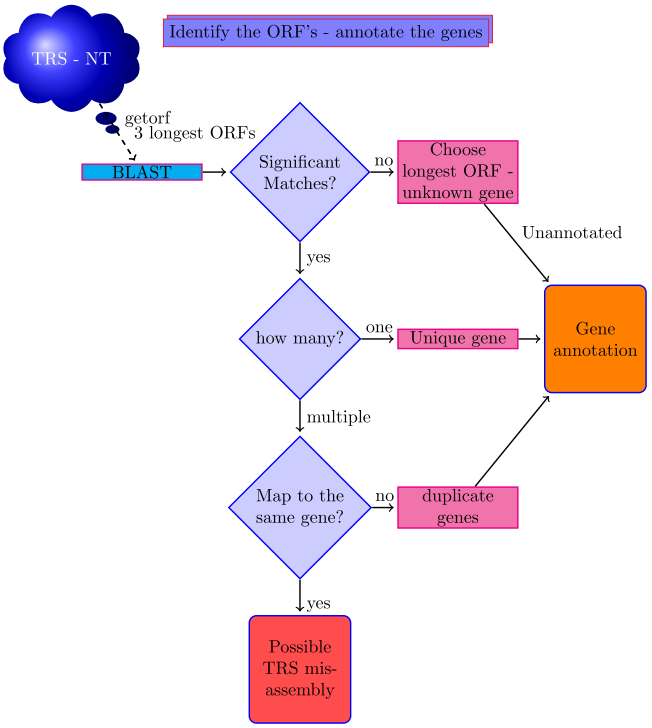

Figure 1. Flowchart for YeATS. For each transcript, the three longest open reading frames (ORF) are obtained using the 'getorf', and these were BLAST'ed to the full non-redundant protein sequences ('nr') database. Based on the number of significant matches, the transcriptome is partitioned. Unique genes have only one significant match, erroneous transcripts have multiple ORFs matching the same gene, while duplicate genes have multiple distinct matches.

3. More than one ORF is less than the cutoff.

   (a) The ORFs map to different fragments of the same protein. This points to an error in the sequencing or the assembly, which breaks down the contiguous ORF into two fragments.

   (b) The ORFs map to different proteins - these are instances of a transcript having two valid ORFs. We duplicate the transcript, associating each one to a different protein sequence.

To produce the uniquely annotated set of genes, we ignored entries with the keywords chromosome, hypothetical, unnamed, unknown and uncharacterized, in order to have a functional characteristic in the annotation, provided the final annotated entry has low E-value. Also, apart from comparing E-values, we also compare the BLAST score, choosing an ORF as unique if its BLAST score was more than twice any other BLAST score, even if other scores satisfied the E-value criteria.

Algorithm 1 describes the process of merging transcripts (SI Figure 1). For a given length (which varies from 5 to 15 in this case), the 5' and 3' sequences and identifiers of each transcript are stored in new string databases: 3'=Begin; 5'=End. Repetitive strings (strings that have only two letters) are ignored, as it is difficult to ensure their uniqueness. For each string of *n* length in the Begin (3') string database, we find whether: a) unique matches of *n* length (one-to-one mapping) are present in the End (5') string database and b) that the prefixes (initial transcript identifiers) of the transcripts are the same.

Algorithm 1. MergeTRS - Merge two transcripts

Input: ϕ_{TRS} ⇐ Set of transcripts
Output: $\phi_{TRSMERGED}$: Pairs of transcripts that can be merged
begin
  $\phi_{TRSMERGED} \leftarrow 0$;
  while *NewStatesAdded* **do**
    foreach TRS_i *in* ϕ_{TRS} **do**
      $\phi_{BEGIN} \leftarrow 0$;
      $\phi_{END} \leftarrow 0$;
      foreach *len:5..15* **do**
        AddBeginingofTRS(ϕ_{BEGIN},TRS_i,*len*);
        AddEndofTRS(ϕ_{END},TRS_i,*len*);
      end
      foreach $string_i$ *in* ϕ_{BEGIN} **do**
        /* ignore strings that have less than 3 letters, these are repetitive*/
        IgnoreRepeats($string_i$);
        if(∃ only one $string_j$ in ϕ_{END}) such that prefixof(TRS_i) == prefixof(TRS_j))[
        $\phi_{TRSMERGED} \leftarrow$
        AddtoMergeableSet(TRS_i,TRS_j);
        ]
      end
    end
  end
  return $\phi_{TRSMERGED}$;
end

Algorithm 2 describes the iterative method for identifying homologous genes in the genome based on the transcriptome. First, the transcriptome is converted to a set of protein sequences by choosing the appropriate ORF (described above) as the representative protein sequence, and a BLAST database (TRSDB) is created. An input protein sequence (possibly from another organism) of a gene of interest is used to query TRSDB using BLAST[29]. This results in a set of significant transcript matches which is pruned based on a cutoff identity (40% in this case) and the criterion that the sequence length should not differ more than another parameterizable value

Algorithm 2. FindGene - Iterative method to identify homologous genes based on the transcriptome

Input: *G* ⇐ Amino acid sequence of gene
Input: *TRSDB* ⇐ BLAST database of the protein sequences from each transcript, choosing the longest ORF as the representative protein sequence
Input: *identitycutoff* ⇐ Ignore matches which are less than *identitycutoff* % identical to the sequence under consideration
Input: *lengthcutoff* ⇐ Ignore matches where the sequence length differs by more than *lengthcutoff* % from the sequence under consideration
Output: ϕ_{genes}
begin
  $\phi_{genes} \leftarrow G$;
  $\phi_{processed} \leftarrow 0$;
  NewStatesAdded ← 1;
  while *NewStatesAdded* **do**
    NewStatesAdded ← 0;
    foreach G_i *in* ϕ_{genes} *such that* G_i *is not in* $\phi_{processed}$ **do**
      $\phi_{processed} \leftarrow G_i$;
      ϕ_i^{BLAST} = BLAST G_i on *TRSDB*;
      foreach TRS_i *in* ϕ_i^{BLAST} **do**
        difflength ←
        length(G_i) – *length*(TRS_i) ;
        if(identity(TRS_i, G_i) > *identitycutoff* ^
        (*difflength* < *lengthcutoff*)) [
        NewStatesAdded ← 1;
        $\phi_{genes} \leftarrow TRS_i$;
        ]
      end
    end
  end
  /* This is not a TRS, but an input - remove this from the set*/
  remove *G* from ϕ_{genes};
  return ϕ_{genes};
end

(50 in this case). Both these transcripts are now potential genes, and the above mentioned process is repeated for each of them, until no new transcripts are added.

The raw counts for each transcript is normalized according to Equation 1, assuming a read length of 100.

$$Score_{normal} = 100 * [Score_{raw}/(Length(transcript))]; \quad (1)$$

The sequence alignment was done using ClustalW[31]. The alignment images were generated using SeaView[32].

The runtimes for most of the processing required in YeATS is a few hours on a simple 16 GB, 16-core machine, barring the search for homologies in the BLAST 'nr' database. This search can be significantly accelerated when the organism under investigation has well-annotated protein databases (as in the current case), much in lines of the newly introduced SMARTBLAST (http://blast.ncbi.nlm.nih.gov/smartblast/), to runtimes under a day.

In vitro methods

Total RNA was isolated from the xylem region immediately external to the heartwood of a 16 year-old black walnut. The tree was felled in November, cross sections about 1 inch thick were taken from the base and dropped immediately into liquid nitrogen. After the sections were fully frozen they were transported to the lab on dry ice. The transition zone was then chiseled and the xylem was ground using a freezer mill. The RNA was extracted from 100g of ground wood using lithium chloride extraction buffer, and subsequently treated with DNAse (to remove genomic DNA) using an RNA/DNA Mini Kit (Qiagen, Valencia, CA) per the manufacturers protocol. Presence of RNA was confirmed by running an aliquot on an Experion Automated Electrophoresis System (Bio-Rad Laboratories, Hercules, CA).

The cDNA libraries were constructed following the Illumina mRNA-sequencing sample preparation protocol (Illumina Inc., San Diego, CA). Final elution was performed with 16 μL RNase-free water. Each library was run as an independent lane on a Genome Analyzer II (Illumina, San Diego, CA) to generate paired-end sequences of 85bp in length from each cDNA library.

Prior to assembly, all reads underwent quality control for paired-end reads and trimming using Sickle[33]. The minimum read length was 45bp with a minimum Sanger quality score of 35. The quality controlled reads of 19 libraries from *J. regia* were *de novo* assembled with Trinity v2.0.6[14] (standard parameters with minimum contig length of 300bp) (manuscript in submission, bioproject id PRJNA232394). Subsequently, the reads from the TZ from *J. nigra* was aligned to this transcriptome and counts obtained by BWA's short read aligner v.0.6.2 ('bwa aln') (http://bio-bwa.sourceforge.net/)[34]. The Illumina reads for the transition wood transcriptome can be accessed at http://www.ncbi.nlm.nih.gov/sra/SRX404331.

Results

The input dataset to the YeATS tool was a set of transcripts, transcript identifiers and their corresponding raw counts (see Supporting information), obtained from the tissue at the heartwood/sapwood transition zone (TZ) in black walnut (*Juglans nigra* L.) (Figure 2). These raw counts were normalized (see Methods), and transcripts with zero counts were ignored (see rawcounts.normalized.TZ in Dataset 1). There were ~24K such transcripts $(\phi^{TZ}_{transcript})$.

Dataset 1. YeATS Dataset

http://dx.doi.org/10.5256/f1000research.6617.d49730

README

FASTADIR.tgz : 24k transcripts

ORFS.tgz : open reading frames from 24k transcripts computed from the 'getorf' tool from the Emboss suite.

list.merged.txt : transcripts that have been merged based on overlapping ends

High.TZ.genome.annotated.csv : transcripts having only one ORF with a high significance match

Lower.TZ.genome.annotated.csv : transcripts having only one ORF with a lower significance match

TZ.genome.annotated.none.csv : transcripts with no match

TZ.genome.errors : transcripts which have two ORFs matching with high significance to the same gene

TZ.genome.annotated.morethanone.csv : transcripts having more than one ORFs which match to different genes with high significance

rawcounts.TZ: Raw counts

rawcounts.normalized.TZ: Normalized counts

In order to associate a transcript to a specific open reading frame (ORF), the ORFs of $\phi^{TZ}_{transcript}$ is obtained using 'getorf' from the Emboss suite[30] (see ORFS.tgz in Supporting information) (Figure 1). The three longest ORFs for each transcript is BLAST'ed

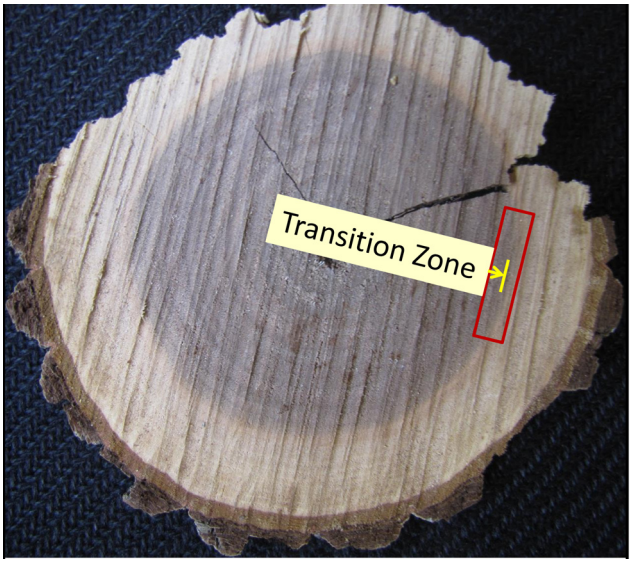

Figure 2. Heartwood/sapwood transition zone in black walnut. A cross section of a mature black walnut (*Juglans nigra*) stem showing the light-colored sapwood (Secondary xylem), darkly colored heartwood which contains no living cells. The transition zone (TZ) is immediately external to the heartwood highlighted by the yellow line in the red box. Cell death is actively occurring in this TZ tissue.

to the full non-redundant protein sequences ('nr') database, and the results were used to characterize the genes.

There were ~1200 transcripts that had possible sequencing or assembly errors, ~22K transcripts that had significant matches (E-value<E-12) in the 'nr' database, 113 transcripts that had lower matches (E-12<E-value<E-08) in the 'nr' database, ~700 transcripts that had no matches in the 'nr' database and about 200 transcripts that could be merged based on overlapping amino acid sequences. We describe these in detail below.

Possible sequencing error or mis-assembly of transcripts

We observed transcripts that had multiple ORFs that matched to the same gene with high significance (E-value<E-10). The possibility that such an occurrence is not an experimental artifact is low. Transcript C15259_G1_I1 is one such example, having two ORFs - ORF_36 (length = 144) and ORF_9 (length = 122), both of which match to the mitochondrial ATP-dependent Clp protease proteolytic subunit 2[35] (GenBank: CAN64666.1) from *Vitis vinifera* with E-values of 6E-92 and 7E-45, respectively. Figure 3 shows the alignment of these two ORFs to the *Vitis vinifera* protein indicated the possible site of the sequencing error or transcript misassembly. This aspect of the YeATS methodology can be used to estimate the sequencing and transcript assembly error rate. For example, in the current transcriptome of the walnut TZ, we found a 5% (1200 out of 24,000) error rate.

Long repeat within the same transcript

A small number of transcripts had long repeats (on the reverse strand), as identified by transcripts that had multiple identical ORFs. For example, transcript C50369_G5_I2 has two ORFs (length = 143) that matched to an uncharacterized protein (Uniprot id: XP_009362671, E-value= 4e-13). These ORFs were located on the reverse strand, and were exactly the same (Figure 4). There were only 8 such cases.

Merging transcripts

About ~200 transcripts have been merged using conservative metrics by YeATS (see Methods, list.merge in Supporting information). For example, transcripts C55368_G1_I3 and C55368_G2_I1 were merged based on a stretch of 12 amino acids (NFDENRGALNSH) (Figure 5). The indicated single nucleotide difference might be the reason for the failure of the assembly program to merge these two

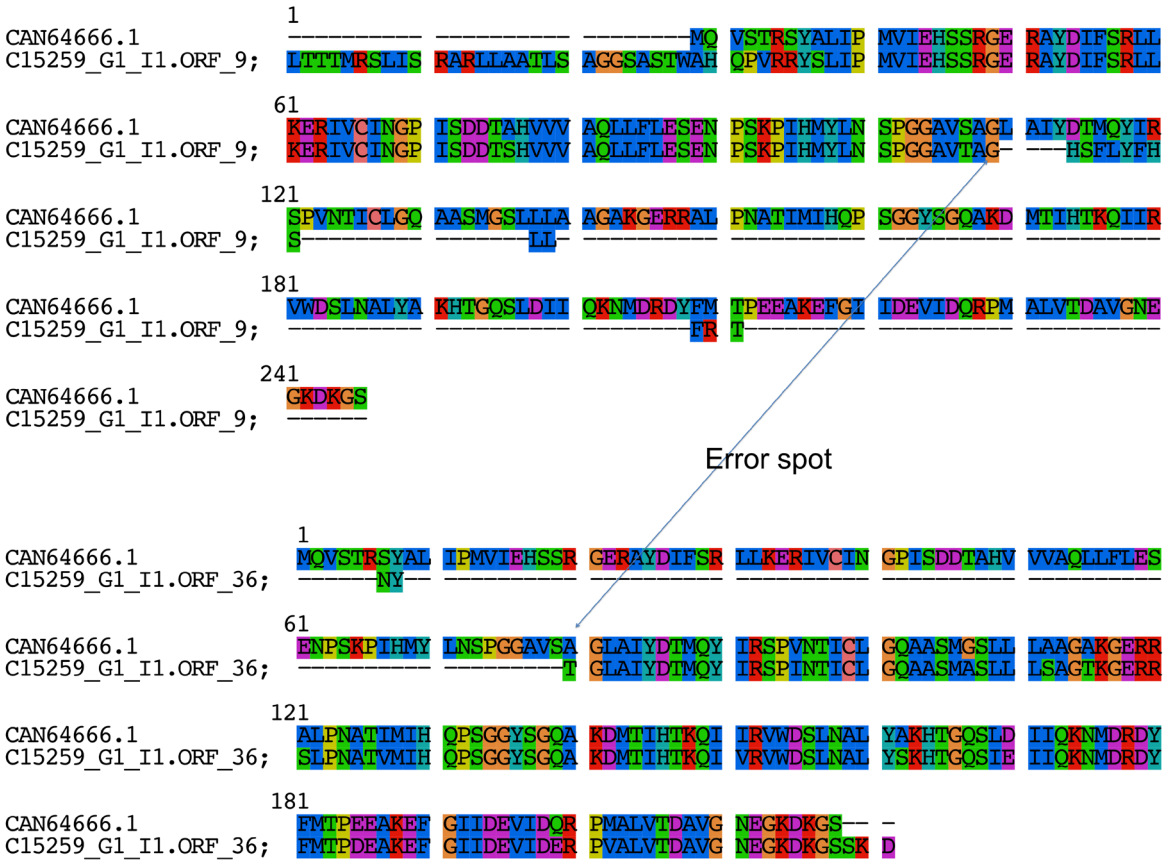

Figure 3. Error detection in sequencing or transcript assembly by YeATS. Transcript C15259_G1_I1 has two ORFs - 9 and 36 - both of which match to the mitochondrial ATP-dependent Clp protease proteolytic subunit 2, mitochondrial (GenBank: CAN64666.1) from *Vitis vinifera* with E-values of 6E-92 and 7E-45, respectively. It is likely that the error occurred near the amino acid sequence 'SAG' marked in the figure. The current transcriptome of the walnut TZ had a 5% (1200 out of 24,000) error rate for this class of error.

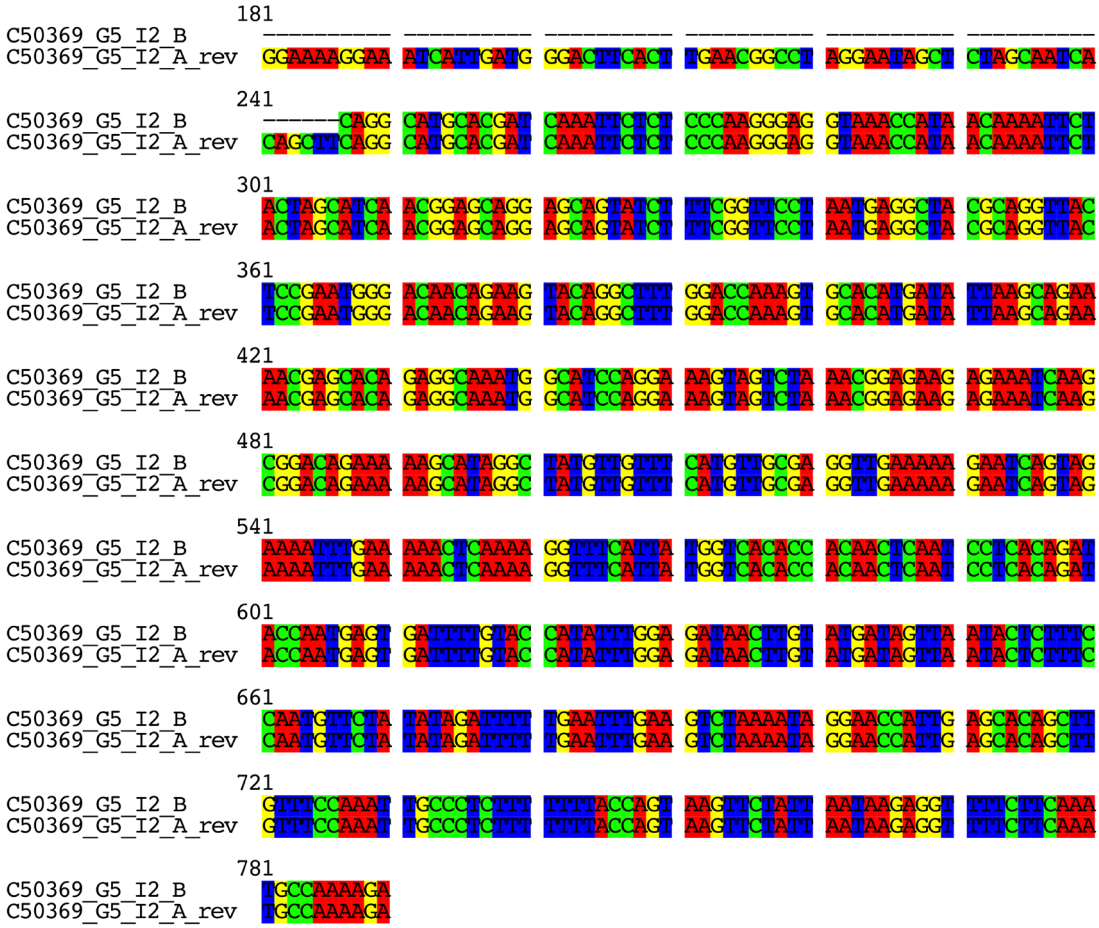

Figure 4. Erroneous transcripts with an exact long repeat (on the reverse strand). Transcript C50369_G5_I2 had an ORF (length = 143, Uniprot id: XP_009362671, uncharacterized protein), with an exact match on the reverse strand. There were only eight such cases, and they could be manually corrected.

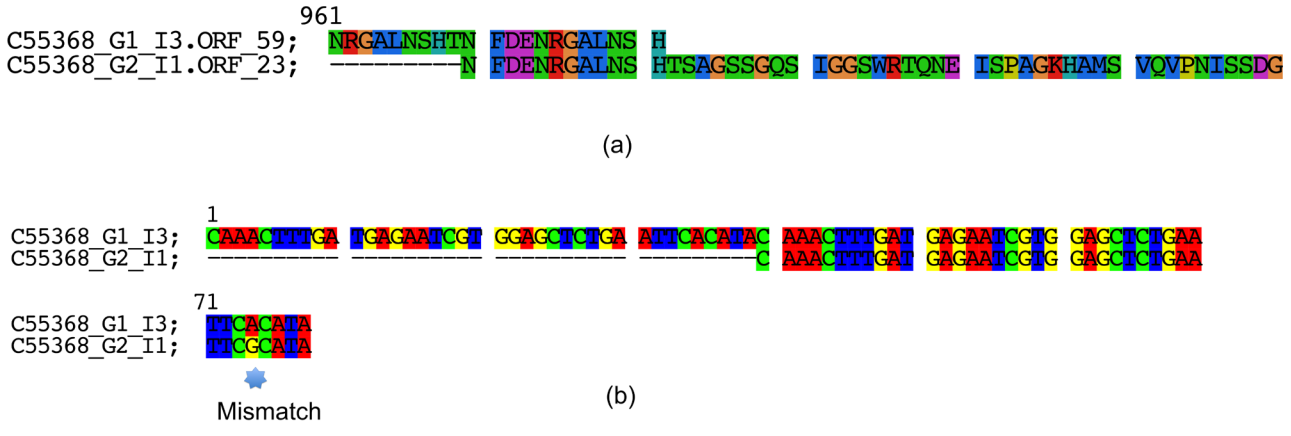

Figure 5. Transcripts that could be merged. (**a**) Transcripts C55368_G1_I3 and C55368_G2_I1 could be merged based on a stretch of 12 amino acids (NFDENRGALNSH) obtained from their ORFs. (**b**) The partial nucleotide sequences of these transcripts shows the repeat with only a single nucleotide difference. The indicated single nucleotide difference may explain the failure of the assembly program to merge these two transcripts. Interestingly, the transcript C55368_G1_I3 had two exact repeats of this stretch at the end which may have contributed to the failure of the assembly program to merge these transcripts.

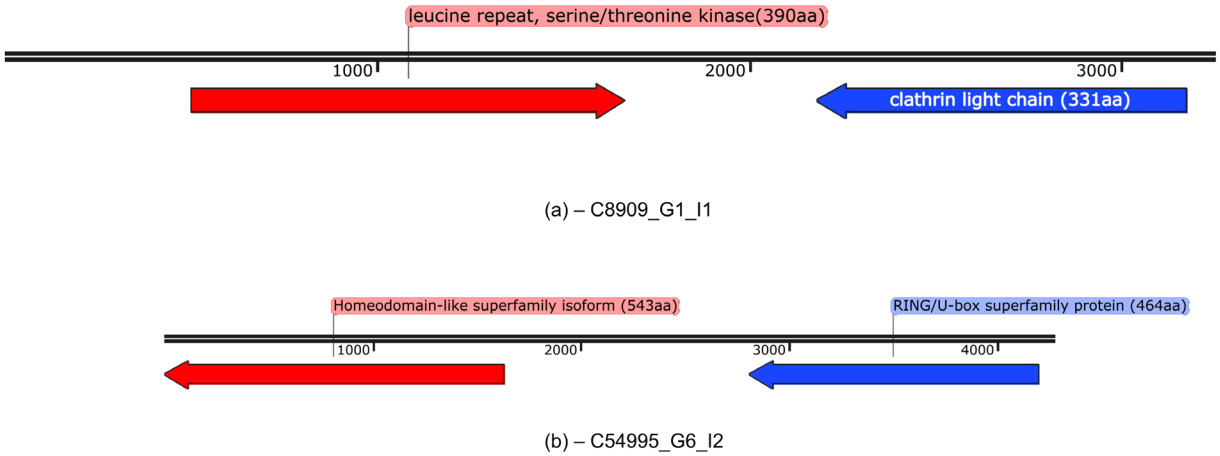

Figure 6. Identification of transcripts encoding multiple genes. These ORFs belong to the same transcript, and have significant matches to different proteins. (**a**) Genes on the reverse strand, having no overlap - clathrin light chain (value=3E-126) and a leucine repeat rich receptor-like serine/threonine protein kinase (E-value=0). (**b**) Genes on the same strand, having no overlap - RING/U-box superfamily protein (E-value=7E-149) and a homeodomain-like superfamily protein isoform (E-value=0).

transcripts. Transcript C55368_G1_I3 had two exact repeats of this stretch, which is a likely assembly error.

Single transcripts with two ORFs

Some transcripts were associated with multiple ORFs with distinct significant matches in the 'nr' database. We demonstrate this for the transcript C8909_G1_I1, which had two ORFs - ORF_104 (length = 331) and ORF_45 (length = 390) which matched to a clathrin light chain[36] (Uniprot id:XP_006481016.1, E-value=3E-126) and a leucine repeat rich receptor-like serine/threonine protein kinase[37] (Uniprot id: XP_007026739.1, E-value=0), respectively. These ORFs were on opposite strands, and did not overlap. It was not possible to ascertain which was the correct gene product, and it is a distinct possibility that both strands were transcribed[38]. A slightly different situation arose when both the ORFs were on the same strand[39], as in the case of the transcript C54995_G6_I2. For example, in transcript C54995_G6_I2, there were two ORFs - ORF_157 (length = 464) and ORF_231 (length = 543) that matched to a RING/U-box superfamily protein[40] (Uniprot id: XP_007042454.1, E-value=7E-149) and a homeodomain-like superfamily protein isoform[41] (Uniprot id: XP_007030696.1, E-value=0), respectively. Both of these proteins were on the same (reverse) strand of the transcript. These transcripts are candidates for chimeric[42] or fusion[43] genes, since the ribosome is known to bypass small nucleotide stretches separating two ORFs[44].

Highly transcribed genes

Table 1 shows the transcripts with the highest counts. Interestingly, the most abundant transcript had no homologous counterpart in the full BLAST 'nr' or 'nt' database (GenBank accession: C52369_G2_I1). A proline-rich protein (PRP), a part of the protein superfamily of cell wall proteins consisting of extensins and nodulins, was found to have the second most abundant transcript[23,45]. Proline comprises 19% of the amino acids in the ORF of this transcript. PRPs are found as structural proteins in wood, and it was hypothesized that these proteins occur in the xylem cell walls during lignification, and influence the properties of wood[46]. PRPs were associated with carrot storage root formation[47], were wound and auxin inducible[47] and implicated in cell elongation[48]. PRPs are also an integral component of saliva responsible for the precipitation of antinutritive and toxic polyphenols by forming complexes[49]. Two DNAJ/HSP40 chaperone proteins, which are involved in proper protein folding, transport and stress response, showed high transcriptional levels[28]. Two DNAJ/HSP40 chaperone homologs (GenBank accession id: BI677935 and BI642398) were shown to be differentially expressed during summer at the sapwood/heartwood TZ of black locust[50]. The transcription levels of dehydrin-related proteins were shown to be seasonally regulated in the wood of deciduous trees[26,51]. However, this dehydrin protein is homologous to a 24kDa dehydrin (Uniprot id: AGC51777) from *Jatropha manihot*, a drought resistant plant[52], unlike the ~100kDa proteins investigated in 26. Senescence-associated proteins, and the related tetraspanins, were also highly transcribed[27]. One highly expressed transcript was homologous to a protein that is yet to be characterized.

Finding genes

We demonstrated the (iterative) gene finding methodology in YeATS on a transcription factor that has an AP2 DNA binding motif (RAP2.6L in *Arabidopsis*, At5g13330)[53]. This protein showed differential tissue specific expression, and is likely to be involved in plant developmental processes and stress response[54]. Recently, the sequence of a homolog of RAP2.6L was deduced (Uniprot id: C1KH72, JnRap2) from an EST sequence isolated from tissue at the heartwood/sapwood TZ in black walnut (*Juglans nigra* L.), and its role in the integration of ethylene and jasmonate signals in the xylem and other tissues was established[55,56]. Using the sequence of JnRap2, we probed for other RAP2 genes in the TZ of walnut. We found three possible genes (C38523_G2_I1, C53728_G7_I1 and C53728_G7_I2) (Figure 7). It was observed that C53728_G7_I2

Table 1. A sample of highly transcribed genes with high normalized counts (NC). There are several highly transcribed genes in the representative sample of the transcriptome from the tissue at the heartwood/sapwood transition zone (TZ) in black walnut that did not have any significant homologs (NSL) in the complete 'nr' or 'nt' database. For the 'nr' database, we use the three longest ORFs as query. The significance of dehydrins, senescence-associated and DNAJ proteins can be observed through their transcription abundance.

ID	NC	Description	E-value
C52369_G2_I1	43040	NSL (putative extensin based on amino acid composition)	-
C51134_G2_I2	15200	ref|XP_008224364.1|PREDICTED: extensin-like [*Prunus mume*]	1e-08
C40830_G1_I1	14169	ref|XP_006365673.1|dnaJ protein homolog isoform X2 [*Solanum tuberosum*]	0
C46581_G1_I1	10651	PREDICTED: Probable zinc transporter protein [*Phoenix dactylifera*]	8e-09
C51134_G2_I3	10631	emb|CAN59948.1|hypothetical protein VITISV_043422 [*Vitis vinifera*]	6e-09
C44353_G2_I1	7769	gb|AGC51777.1|dehydrin protein [*Manihot esculenta*]	6e-09
C44353_G1_I1	6652	gb|AAF01465.2|AF190474_1 bdn1 [*Paraboea crassifolia*]	2e-19
C43130_G3_I1	6601	gb|KEH16988.1|senescence-associated protein, putative [*Medicago truncatula*]	2e-129
C44922_G1_I1	5584	ref|XP_008363477.1|tetraspanin-3-like [*Malus domestica*]	2e-169
C40830_G1_I2	5113	ref|XP_007010484.1|DNAJ [*Theobroma cacao*]	0

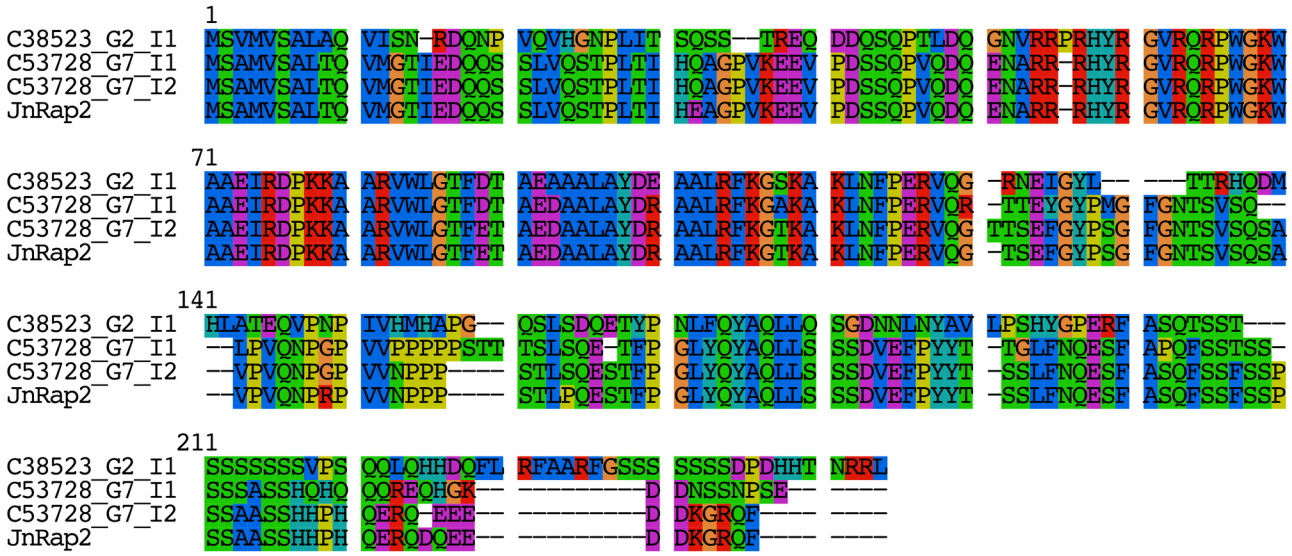

Figure 7. Finding genes from a template sequence. Multiple sequence alignment of possible genes for a transcription factor that had a AP2 DNA binding motif compared to JnRap2, which was deduced from an EST sequence obtained from tissue at the heartwood/sapwood transition zone in black walnut.

was closest to the JnRap2 gene (97.4% identity, 98.2% similar), and is probably the same gene. C53728_G2_I1 was also significantly homologous to the JnRap2 gene (84.4% identity, 92.4% similar), and it appears to be an allelic or splice variant, a conflict that can be resolved after the publication of the complete walnut genome. Raw counts (see Supporting information) demonstrated that the transcript C38523_G2_I1 had negligible expression levels in TZ, corroborating the previous detection of only one RAP2 protein in 55.

Transcripts with no significant matches in the 'nr' database - possible long non-coding RNA genes?

The top three ORFs of ~600 transcripts had no match in the BLAST 'nr' database. Although these may be unique genes, another possibility that must be considered is that these are non-coding RNA genes[2]. The nucleotide sequences of these 600 transcripts were BLAST'ed to the database of noncoding RNAs in *Arabidopsis*[22]. Three matches were identified: C52424_G5_I11, C52424_G5_I4 and C53565_G3_I1.

Both C52424_G5_I11 and C52424_G5_I4 are homologous to CR20, a cytokinin-repressed gene in excised cotyledons of cucumber, hypothesized to be non-coding RNA[57]. Analogous to the current work, the CR20 gene had alternate splicing[57]. C53565_G3_I1 had a 100% match to the *Arabidopsis* locus ATMG01380, a mitochondrial 5S ribosomal RNA, which is a component of the 50S large subunit of mitochondrial ribosome[58].

Discussion

High-throughput mRNA sequencing (RNA-Seq) has revolutionized the field of transcript discovery, providing several advantages over traditional methods[7,8]. Following isolation and fragmentation of RNA and subsequent generation of cDNA libraries, a high-throughput sequencing platform is selected to generate short reads[59]. Reconstruction of transcripts from these short reads (assembly) may be performed using a reference genome or *de novo* algorithms[15–18,21,60]. Sequencing biases, variable coverage, sequencing errors, alternate splicing and repeat sequences are some of the challenges faced by these assemblers[14,61].

Several post assembly computational tools provide further curation of transcripts resulting from the assemblers. The curation step involves identifying redundancies[19,20], finding coding regions[62], annotating the transcripts (https://transdecoder.github.io/) and detecting inaccuracies by aligning the transcripts to the genome[63]. In the current work, we present an integrated workflow for RNA-seq analysis (YeATS). YeATS includes most features of the tools mentioned above. Additionally, YeATS delivers several capabilities absent in these tools. A comprehensive BLAST analysis of the top three open reading frames of each transcript enables the identification of erroneous transcripts arising out of sequencing or assembly errors. These erroneous transcripts can be classified as: a) transcripts that have not been merged, b) transcripts that result in broken ORFs and c) transcripts that have long improbable repeats. Finally, YeATS provides annotation of the genes, enumerates homologous genes based on a template sequence and specified similarity threshold and identifies transcripts with multiple ORFs. The ribosome is known to bypass small nucleotide stretches separating two ORFs[44]. These are rare events, however, and thus unlikely to apply to the ~1200 transcripts that have broken ORFs pointing to the same gene[64]. Transcripts having multiple ORFs on the same strand are good candidates for chimeric[42] or fusion[43] genes dependent on ribosome bypassing.

The current work reveals and corroborates several aspects of the biology of hardwood trees. Probably, the most interesting is the detection of a highly transcribed gene (C52369_G2_I1) with no known homologs in the complete protein and nucleotide BLAST database, or significant matches in a database of long non-coding RNA genes[22]. If indeed the longest ORF of this transcript encodes a protein, it is 143 amino acids long, and is leucine (18%), histidine (13%) and valine (10%) rich (Figure 8). Although it is likely that this is a protein with leucine rich repeats, these proteins are typically larger proteins[65]. On the other hand, histidine and valine rich extensins have been reported to be constituents of plant cell walls of dicots[23]. The regulatory stimuli of extensins are different for monocots (which also have different amino acid composition) and dicots[23]. A significant presence of extensin-like proteins in the cell wall of both developing and mature xylem (wood) have been reported for pine[46,66]. The publication of the walnut genome will aid the characterization of these genes by elucidating its promoter sequences.

Well characterized proteins like proline-rich proteins[25,46], dehydrins[26], senescence-associated proteins[27] and DNAJ/HSP40 chaperone[50] proteins were also abundant in the transcriptome. While *Arabidopsis* supports secondary growth, it fails to accumulate wood; it is therefore interesting to identify highly transcribed genes that are missing in the *Arabidopsis* proteome (Table 2). The

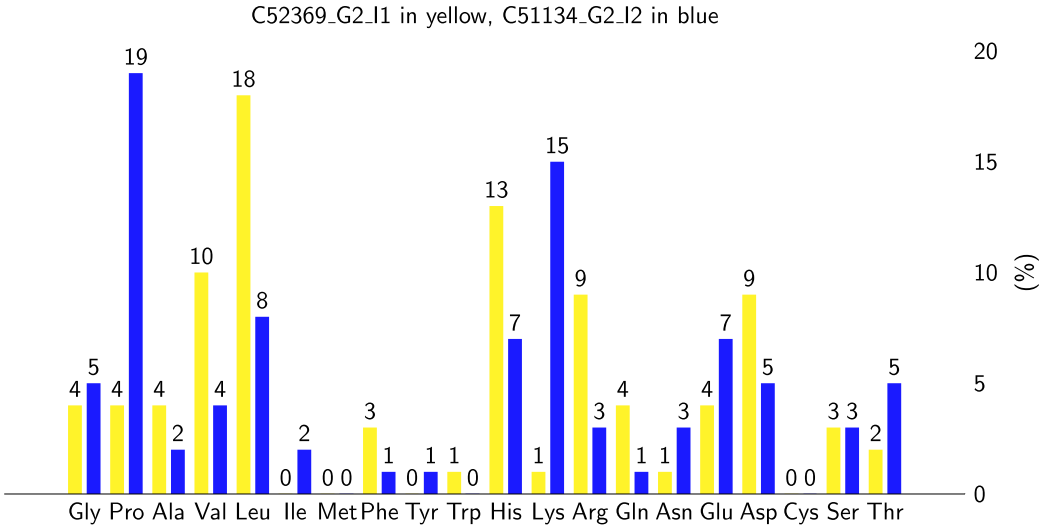

Figure 8. Percentage amino acid composition of the two most highly transcribed genes. C52369_G2_I1 has a high percentage of leucine, histidine and valine, and is a putative extensin. C51134_G2_I2 is proline and lysine rich, and is homologous to an extensin and nodulin.

Table 2. Identifying highly transcribed genes that are not present in the *Arabidopsis* proteome. The wood quality of walnut and *Arabidopsis* are quite different. It is informative to identify genes (proteins) that are absent in *Arabidopsis*, since they are likely to be responsible for the differences. The DNAJ/HSP40 chaperone, dehydrins and tetraspanin proteins are found in the *Arabidopsis* proteome, while the putative extensin, the proline-rich protein, a probable zinc transporter protein, an uncharacterized protein and senescence-associated protein appear to be unique to the walnut proteome.

TRS	Arabidopsis Id	Description	E-value	Significant?
C52369_G2_I1	AT5G04990.1	SUN1, ATSUN1 | SAD1/UNC-84 domain protein	0.75	
C51134_G2_I2	AT3G18440.1	AtALMT9, ALMT9 | aluminum-activated malat	0.046	
C40830_G1_I1	AT5G22060.1	ATJ2, J2 | DNAJ homologue 2 | chr5:730379	0	Y
C46581_G1_I1	AT5G51930.1	Glucose-methanol-choline (GMC) oxidore	8.1	
C51134_G2_I3	AT1G79090.2	FUNCTIONS IN: molecular function unkno	1.3	
C44353_G2_I1	AT1G76180.2	ERD14 Dehydrin family protein | chr1:28	1e-05	Y
C44353_G1_I1	AT1G20450.2	LTI29, LTI45, ERD10 | Dehydrin family pro	1e-07	Y
C43130_G3_I1	AT1G72110.1	O-acyltransferase (WSD1-like) family p	1.7	
C44922_G1_I1	AT3G45600.1	TET3 | tetraspanin3 | chr3:16733973–16735	8e-156	Y
C40830_G1_I2	AT3G44110.1	ATJ3, ATJ | DNAJ homologue 3 | chr3:15869	1e-179	Y

DNAJ/HSP40 chaperone, dehydrins and tetraspanin proteins are found in the *Arabidopsis* proteome (TAIR10_pep_20101214[67]), while the putative extensin, the proline-rich protein, a probable zinc transporter protein, an uncharacterized protein and senescence-associated protein appear to be unique to the walnut proteome.

Also, we corroborated the presence of a transcription factor that has a AP2 DNA binding motif[53,55], and identify additional splice/allelic variants with similar transcriptional levels. Once again, the knowledge of the walnut genome would enable a more profound understanding of such genes.

Conclusions

In summary, the current work elucidates an integrated workflow for RNA-seq analysis with several innovative features for identifying and correcting erroneously assembled transcripts. We demonstrated this workflow by characterizing the transcriptome of the tissue at the heartwood/sapwood TZ in black walnut.

Data availability

F1000Research: Dataset 1. YeATS Dataset, 10.5256/f1000research.6617.d49730[68]

Software availability

Latest source code

https://github.com/sanchak/YEATSCODE2

Archived source code as at the time of publication

http://dx.doi.org/10.5281/zenodo.33137

Software license

GNU General Public License version 3.0 (GPLv3)

Author contributions

The YeATS tool suite was designed by Chakraborty, Britton and Wegrzyn did the analysis of the transcriptome, Woeste isolated the RNA, Butterfield was involved in the validation. The rest of the authors were involved in various aspects of the study design. Chakraborty wrote the first draft and the rest of the authors were involved in the editing.

Competing interests

No competing interests were disclosed.

Grant information

The authors wish to acknowledge support from the California Walnut Board and UC Discovery program.

I confirm that the funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Acknowledgements

We are grateful to Mary Lou Mendum for her inputs in preparing the manuscript.

Supplementary figure

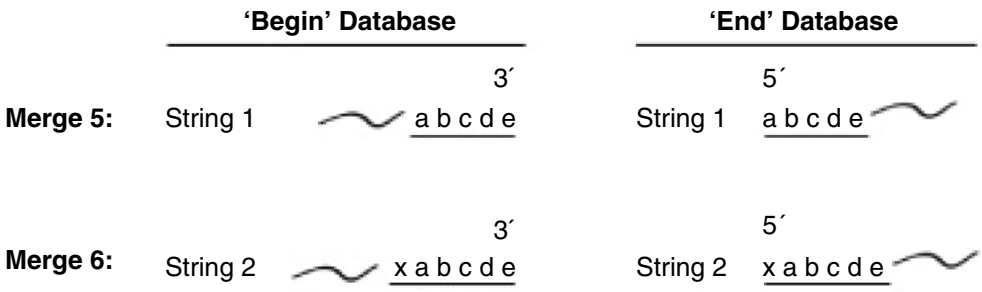

Figure S1. Pictoral depiction of 'Begin' and 'End' database. The contents of the 'Begin' and 'End' databases are represented along with the shared amino acid residues at the 3' and 5' ends, respectively. The 'merge5' command identified the String1 pair in the 'Begin' and 'End' databases due to the shared sequence 'abcde'. The 'merge5' command failed to identify the String2 pair since six residues are shared. The 'merge6' command, however, will recognize String2 in the 'Begin' and 'End' databases, but would fail to recognize pairs that shared seven or more residues at the 3' and 5' 'End'.

References

1. Crick F: **Central dogma of molecular biology.** *Nature.* 1970; **227**(5258): 561–563.
PubMed Abstract | Publisher Full Text
2. Mattick JS, Makunin IV: **Non-coding RNA.** *Hum Mol Genet.* 2006; **15**(Spec No 1): R17–R29.
PubMed Abstract | Publisher Full Text
3. Kakumanu A, Ambavaram MM, Klumas C, *et al.*: **Effects of drought on gene expression in maize reproductive and leaf meristem tissue revealed by RNA-seq.** *Plant Physiol.* 2012; **160**(2): 846–867.
PubMed Abstract | Publisher Full Text | Free Full Text
4. Costa V, Aprile M, Esposito R, *et al.*: **RNA-Seq and human complex diseases: recent accomplishments and future perspectives.** *Eur J Hum Genet.* 2013; **21**(2): 134–142.
PubMed Abstract | Publisher Full Text | Free Full Text
5. Clark TA, Sugnet CW, Ares M Jr: **Genomewide analysis of mRNA processing in yeast using splicing-specific microarrays.** *Science.* 2002; **296**(5569): 907–910.
PubMed Abstract | Publisher Full Text
6. Kodzius R, Kojima M, Nishiyori H, *et al.*: **CAGE: cap analysis of gene expression.** *Nat Methods.* 2006; **3**(3): 211–222.
PubMed Abstract | Publisher Full Text
7. Wang Z, Gerstein M, Snyder M: **RNA-seq: a revolutionary tool for transcriptomics.** *Nat Rev Genet.* 2009; **10**(1): 57–63.
PubMed Abstract | Publisher Full Text | Free Full Text
8. Flintoft L: **Transcriptomics: digging deep with RNA-seq.** *Nat Rev Genet.* 2008; **9**(8): 568.
Publisher Full Text
9. Trapnell C, Pachter L, Salzberg SL: **TopHat: discovering splice junctions with RNA-seq.** *Bioinformatics.* 2009; **25**(9): 1105–1111.
PubMed Abstract | Publisher Full Text | Free Full Text
10. Trapnell C, Roberts A, Goff L, *et al.*: **Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks.** *Nat Protoc.* 2012; **7**(3): 562–578.
PubMed Abstract | Publisher Full Text | Free Full Text
11. Wang L, Feng Z, Wang X, *et al.*: **DEGseq: an R package for identifying differentially expressed genes from RNA-seq data.** *Bioinformatics.* 2010; **26**(1): 136–138.
PubMed Abstract | Publisher Full Text
12. Lohse M, Bolger AM, Nagel A, *et al.*: **Robi*NA*: a user-friendly, integrated software solution for RNA-seq-based transcriptomics.** *Nucleic Acids Res.* 2012; **40**(Web Server issue): W622-7.
PubMed Abstract | Publisher Full Text | Free Full Text
13. Chang Z, Li G, Liu J, *et al.*: **Bridger: a new framework for *de novo* transcriptome assembly using RNA-seq data.** *Genome Biol.* 2015; **16**(1): 30.
PubMed Abstract | Publisher Full Text | Free Full Text
14. Grabherr MG, Haas BJ, Yassour M, *et al.*: **Full-length transcriptome assembly from RNA-Seq data without a reference genome.** *Nat Biotechnol.* 2011; **29**(7): 644–652.
PubMed Abstract | Publisher Full Text | Free Full Text
15. Chu HT, Hsiao WW, Chen JC, *et al.*: **EBARDenovo: highly accurate *de novo* assembly of RNA-seq with efficient chimera-detection.** *Bioinformatics.* 2013; **29**(8): 1004–1010.
PubMed Abstract | Publisher Full Text
16. Schulz MH, Zerbino DR, Vingron M, *et al.*: ***Oases*: robust *de novo* RNA-seq assembly across the dynamic range of expression levels.** *Bioinformatics.* 2012; **28**(8): 1086–1092.
PubMed Abstract | Publisher Full Text | Free Full Text
17. Chang Z, Li G, Liu J, *et al.*: **Bridger: a new framework for *de novo* transcriptome assembly using RNA-seq data.** *Genome Biol.* 2015; **16**(1): 30.
PubMed Abstract | Publisher Full Text | Free Full Text
18. Simpson JT, Wong K, Jackman JD, *et al.*: **ABySS: a parallel assembler for short read sequence data.** *Genome Res.* 2009; **19**(6): 1117–1123.
PubMed Abstract | Publisher Full Text | Free Full Text
19. Fu L, Niu B, Zhu Z, *et al.*: **CD-HIT: accelerated for clustering the next-generation sequencing data.** *Bioinformatics.* 2012; **28**(23): 3150–3152.
PubMed Abstract | Publisher Full Text | Free Full Text
20. Mbandi SK, Hesse U, van Heusden P, *et al.*: **Inferring *bona fide* transfrags in RNA-seq derived-transcriptome assemblies of non-model organisms.** *BMC Bioinformatics.* 2015; **16**(1): 58.
PubMed Abstract | Publisher Full Text | Free Full Text
21. Zerbino DR, Birney E: **Velvet: algorithms for *de novo* short read assembly using de Bruijn graphs.** *Genome Res.* 2008; **18**(5): 821–829.
PubMed Abstract | Publisher Full Text | Free Full Text
22. Xie C, Yuan J, Li H, *et al.*: **NONCODEv4: exploring the world of long non-coding RNA genes.** *Nucleic Acids Res.* 2014; **42**(Database issue): D98–D103.
PubMed Abstract | Publisher Full Text | Free Full Text
23. Showalter AM: **Structure and function of plant cell wall proteins.** *Plant Cell.* 1993; **5**(1): 9–23.
PubMed Abstract | Publisher Full Text | Free Full Text
24. Plomion C, Leprovost G, Stokes A: **Wood formation in trees.** *Plant Physiol.* 2001; **127**(4): 1513–1523.
PubMed Abstract | Publisher Full Text | Free Full Text

25. Williamson MP: **The structure and function of proline-rich regions in proteins.** *Biochem J.* 1994; **297**(Pt 2): 249–60.
PubMed Abstract | Free Full Text

26. Sauter JJ, Westphal S, Wisniewski M: **Immunological identification of dehydrin-related proteins in the wood of five species of *Populus* and in *Salix* caprea L.** *J Plant Physiol.* 1999; **154**(5-6): 781–788.
Publisher Full Text

27. Olmos E, Reiss B, Dekker K: **The *ekeko* mutant demonstrates a role for tetraspanin-like protein in plant development.** *Biochem Biophys Res Commun.* 2003; **310**(4): 1054–1061.
PubMed Abstract | Publisher Full Text

28. Bekh-Ochir D, Shimada S, Yamagami A, *et al.*: **A novel mitochondrial DnaJ/Hsp40 family protein BIL2 promotes plant growth and resistance against environmental stress in brassinosteroid signaling.** *Planta.* 2013; **237**(6): 1509–1525.
PubMed Abstract | Publisher Full Text | Free Full Text

29. Camacho C, Madden T, Ma N, *et al.*: **BLAST Command Line Applications User Manual**. 2013.
Reference Source

30. Rice P, Longden I, Bleasby A: **EMBOSS: the European Molecular Biology Open Software Suite.** *Trends Genet.* 2000; **16**(6): 276–277.
PubMed Abstract | Publisher Full Text

31. Larkin MA, Blackshields G, Brown NP, *et al.*: **Clustal W and Clustal X version 2.0.** *Bioinformatics.* 2007; **23**(21): 2947–2948.
PubMed Abstract | Publisher Full Text

32. Gouy M, Guindon S, Gascuel O: **SeaView version 4: A multiplatform graphical user interface for sequence alignment and phylogenetic tree building.** *Mol Biol Evol.* 2010; **27**(2): 221–224.
PubMed Abstract | Publisher Full Text

33. Joshi NA, Fass JN: **Sickle: A sliding-window, adaptive, quality-based trimming tool for fastq files (version 1.33)[software],** 2011.
Reference Source

34. Li H, Durbin R: **Fast and accurate short read alignment with Burrows-Wheeler transform.** *Bioinformatics.* 2009; **25**(14): 1754–1760.
PubMed Abstract | Publisher Full Text | Free Full Text

35. Halperin T, Zheng B, Itzhaki H, *et al.*: **Plant mitochondria contain proteolytic and regulatory subunits of the ATP-dependent Clp protease.** *Plant Mol Biol.* 2001; **45**(4): 461–468.
PubMed Abstract | Publisher Full Text

36. Konopka CA, Backues SK, Bednarek SY: **Dynamics of *Arabidopsis* dynamin-related protein 1C and a clathrin light chain at the plasma membrane.** *Plant Cell.* 2008; **20**(5): 1363–1380.
PubMed Abstract | Publisher Full Text | Free Full Text

37. Afzal AJ, Wood AJ, Lightfoot DA: **Plant receptor-like serine threonine kinases: roles in signaling and plant defense.** *Mol Plant Microbe Interact.* 2008; **21**(5): 507–517.
PubMed Abstract | Publisher Full Text

38. Geiduschek EP: **An introduction to transcription and gene regulation.** *J Biol Chem.* 2010; **285**(34): 25885–25892.
PubMed Abstract | Publisher Full Text | Free Full Text

39. Borthakur D, Basche M, Buikema WJ, *et al.*: **Expression, nucleotide sequence and mutational analysis of two open reading frames in the *nif* gene region of *Anabaena* sp. strain PCC7120.** *Mol Gen Genet.* 1990; **221**(2): 227–234.
PubMed Abstract | Publisher Full Text

40. Deshaies RJ, Joazeiro CA: **RING domain E3 ubiquitin ligases.** *Annu Rev Biochem.* 2009; **78**: 399–434.
PubMed Abstract | Publisher Full Text

41. Dubos C, Stracke R, Grotewold E, *et al.*: **MYB transcription factors in *Arabidopsis*.** *Trends Plant Sci.* 2010; **15**(10): 573–581.
PubMed Abstract | Publisher Full Text

42. Fromm ME, Morrish F, Armstrong C, *et al.*: **Inheritance and expression of chimeric genes in the progeny of transgenic maize plants.** *Biotechnology (N Y).* 1990; **8**(9): 833–839.
PubMed Abstract | Publisher Full Text

43. Mitelman F, Johansson B, Mertens F: **The impact of translocations and gene fusions on cancer causation.** *Nat Rev Cancer.* 2007; **7**(4): 233–245.
PubMed Abstract | Publisher Full Text

44. Gallant J, Bonthuis P, Lindsley D: **Evidence that the bypassing ribosome travels through the coding gap.** *Proc Natl Acad Sci U S A.* 2003; **100**(23): 13430–13435.
PubMed Abstract | Publisher Full Text | Free Full Text

45. Kieliszewski MJ, Lamport DT: **Extensin: repetitive motifs, functional sites, post-translational codes, and phylogeny.** *Plant J.* 1994; **5**(2): 157–172.
PubMed Abstract | Publisher Full Text

46. Bao W, O'Malley DM, Sederoff RR: **Wood contains a cell-wall structural protein.** *Proc Natl Acad Sci U S A.* 1992; **89**(14): 6604–6608.
PubMed Abstract | Free Full Text

47. Ebener W, Fowler TJ, Suzuki H, *et al.*: **Expression of DcPRP1 is linked to carrot storage root formation and is induced by wounding and auxin treatment.** *Plant Physiol.* 1993; **101**(1): 259–265.
PubMed Abstract | Publisher Full Text | Free Full Text

48. Dvoráková L, Srba M, Opatrny Z, *et al.*: **Hybrid proline-rich proteins: novel players in plant cell elongation?** *Ann Bot.* 2012; **109**(2): 453–462.
PubMed Abstract | Publisher Full Text | Free Full Text

49. Baxter NJ, Lilley TH, Haslam E, *et al.*: **Multiple interactions between polyphenols and a salivary proline-rich protein repeat result in complexation and precipitation.** *Biochemistry.* 1997; **36**(18): 5566–5577.
PubMed Abstract | Publisher Full Text

50. Yang J, Kamdem DP, Keathley DE, *et al.*: **Seasonal changes in gene expression at the sapwood-heartwood transition zone of black locust (*Robinia pseudoacacia*) revealed by cDNA microarray analysis.** *Tree Physiol.* 2004; **24**(4): 461–474.
PubMed Abstract | Publisher Full Text

51. Bassett CL, Wisniewski ME, Artlip TS, *et al.*: **Comparative expression and transcript initiation of three peach dehydrin genes.** *Planta.* 2009; **230**(1): 107–118.
PubMed Abstract | Publisher Full Text

52. Maes WH, Achtena WMJ, Reubens B, *et al.*: **Plant–water relationships and growth strategies of *Jatropha curcas* L. seedlings under different levels of drought stress.** *Journal of Arid Environments.* 2009; **73**(10): 877–884.
Publisher Full Text

53. Okamuro JK, Caster B, Villarroel R, *et al.*: **The AP2 domain of *APETALA2* defines a large new family of DNA binding proteins in *Arabidopsis*.** *Proc Natl Acad Sci U S A.* 1997; **94**(13): 7076–7081.
PubMed Abstract | Publisher Full Text | Free Full Text

54. Krishnaswamy S, Verma S, Rahman MH, *et al.*: **Functional characterization of four APETALA2-family genes (*RAP2.6, RAP2.6L, DREB19 and DREB26*) in *Arabidopsis*.** *Plant Mol Biol.* 2011; **75**(1–2): 107–127.
PubMed Abstract | Publisher Full Text

55. Huang Z, Zhao P, Medina J, *et al.*: **Roles of *JnRAP2.6-like* from the transition zone of black walnut in hormone signaling.** *PLoS One.* 2013; **8**(11): e75857.
PubMed Abstract | Publisher Full Text | Free Full Text

56. Huang Z, Tsai CJ, Harding SA, *et al.*: **A cross-species transcriptional profile analysis of heartwood formation in black walnut.** *Plant Mol Biol Report.* 2010; **28**(2): 222–230.
Publisher Full Text

57. Teramoto H, Toyama T, Takeba G, *et al.*: **Noncoding RNA for *CR20*, a cytokinin-repressed gene of cucumber.** *Plant Mol Biol.* 1996; **32**(5): 797–808.
PubMed Abstract | Publisher Full Text

58. Barciszewska MZ, Szymanski M, Erdmann VA, *et al.*: **Structure and functions of 5s rRNA.** *Acta Biochim Pol.* 2001; **48**(1): 191–198.
PubMed Abstract

59. Mardis ER: **The impact of next-generation sequencing technology on genetics.** *Trends Genet.* 2008; **24**(3): 133–141.
PubMed Abstract | Publisher Full Text

60. Haas BJ, Papanicolaou A, Yassour M, *et al.*: ***De novo* transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis.** *Nat Protoc.* 2013; **8**(8): 1494–1512.
PubMed Abstract | Publisher Full Text | Free Full Text

61. Roberts A, Trapnell C, Donaghey J, *et al.*: **Improving RNA-seq expression estimates by correcting for fragment bias.** *Genome Biol.* 2011; **12**(3): R22.
PubMed Abstract | Publisher Full Text | Free Full Text

62. Arrial RT, Togawa RC, Brigido Mde M: **Screening non-coding RNAs in transcriptomes from neglected species using PORTRAIT: case study of the pathogenic fungus *paracoccidioides brasiliensis*.** *BMC Bioinformatics.* 2009; **10**: 239.
PubMed Abstract | Publisher Full Text | Free Full Text

63. Zhao QY, Wang Y, Kong YM, *et al.*: **Optimizing *de novo* transcriptome assembly from short-read RNA-seq data: a comparative study.** *BMC Bioinformatics.* 2011; **12**(Suppl 14): S2.
PubMed Abstract | Publisher Full Text | Free Full Text

64. Herr AJ, Gesteland RF, Atkins JF: **One protein from two open reading frames: mechanism of a 50 nt translational bypass.** *EMBO J.* 2000; **19**(11): 2671–2680.
PubMed Abstract | Publisher Full Text | Free Full Text

65. Jones DA, Jones JDG: **The role of leucine-rich repeat proteins in plant defences.** *Advances in botanical research.* 1997; **24**: 89–167.
Publisher Full Text

66. Allona I, Quinn M, Shoop E, *et al.*: **Analysis of xylem formation in pine by cDNA sequencing.** *Proc Natl Acad Sci U S A.* 1998; **95**(16): 9693–9698.
PubMed Abstract | Publisher Full Text | Free Full Text

67. Lamesch P, Berardini TZ, Li D, *et al.*: **The *Arabidopsis* Information Resource (TAIR): improved gene annotation and new tools.** *Nucleic Acids Res.* 2012; **40**(Database issue): D1202–D1210.
PubMed Abstract | Publisher Full Text | Free Full Text

68. Chakraborty S, Dandekar A, Rao BJ, *et al.*: **Dataset 1 in: YeATS - a tool suite for analyzing RNA-seq derived transcriptome identifies a highly transcribed putative extensin in heartwood/sapwood transition zone in black walnut.** *F1000Research.* 2015.
Data Source

Open Peer Review

Current Referee Status: ✓ ✓ ✓

Version 2

Referee Report 19 February 2016

doi:10.5256/f1000research.7788.r12065

Binay Panda
Genomics Applications and Informatics Technology laboratories (GANIT Labs), Bio-IT Centre, Institute of Bioinformatics and Applied Biotechnology(IBAB), Bangalore, Karnataka, India

Chakraborty et al. implemented a workflow for error estimation and correction, functional annotation and abundance estimation in RNA-seq data. They explored a methodology of analyzing longest ORFs of transcripts, using BLAST, as means to identify important genes. Although BLAST has been very commonly used for annotation, the authors proposed a very systematic approach of dividing the annotations into four sets based on the quality of the ORFs and their functional assignments.

Overall comments:
The authors have done a good deal of work in exploring an important topic. The article gives many ideas worth exploring and directions for annotating data from *de novo* transcriptome sequencing. However, I suggest that the authors pay special attention to the usage of different terms (genome, transcriptome, RNA-seq, reads etc.) and be consistent in their usage throughout the manuscript. Additionally, the figure legends need complete re-writing and the perl scripts need to be included in the supporting information. Several explanations need to be provided throughout the manuscript (details below).

A single round of proofreading will hugely improve the manuscript.

Specific suggestions/questions:
The title states that YeATS identifies a highly transcribed putative extension. This is a bit misleading. YeATS takes as input already assembled transcripts from Trinity, estimates their expression, employs BLAST to try and assign a function to the most highly transcribed gene, and finds no known homologs. Only from the manual examination of the amino acid content of its longest ORF do the authors come to the conclusion that it is a putative extension. It will perhaps be better to mention the error-detection and estimation capabilities of YeATS as its strongest points.

Introduction: What evidence is there to suggest that a putative protein with a high percentage of leucine, histidine and valine is a probable extension?

Where does Algorithm 1 fit in? Why is it needed? Merging of transcripts is mentioned for the first time in methods, without an explanation of why it is important and where are its potential applications? The authors should explain the algorithm and how it serves in detecting error in assembly/sequencing, and what kind of transcripts should be used in merging.

Algo 1 - Why is the length range defined as 5 to 15? Is there an explanation behind the selection? Also,

5-15 is nucleotides or amino acids?

Algo 1 - Please clarify why 2-letter strings have been called a 'repetitive'? They should be ignored, agreeably, because they are too short to give any reliable results, but it is misleading to call them repetitive?

In algorithm 1, what is meant by 'prefix' of transcripts? If the workflow, YeATS, is so strictly dependent on the Trinity transcript headers, it needs to be clearly mentioned in the article.

Is 'TRS' equivalent to 'transcript'? Please include a list of abbreviations used in the manuscript at the beginning of the manuscript.

Is there a rationale behind running BLAST a second time, when from the very first BLAST run, which gives the best ORF selection, the gene functional annotation can also be obtained? This could make use of the Algorithm 2 and reduce the overall runtime.

Algo2 - The genome does not picture anywhere here, therefore, is misleading to say 'identifying homologous genes in the genome based on the transcriptome.'

Algo2 - is 'lengthcutoff' a % or number of nucleotides or amino acids? 'Input' says %, whereas in the algo it is simple difference.

Algo2 - 'Both these transcripts are now potential genes', which 2 transcripts are the authors talking about?

Equation 1 - Which raw counts are these? How are they obtained?

What is 'score' in equation 1?

Sequence alignment of what was done using ClustalW? What was it used for?

In vitro methods
The authors should list the 19 different samples whose cDNA libraries were sequenced. Were these combined and assembled as a single transcriptome? Is this used a reference for read alignment and counts estimation? If not, what is the reference for read count estimation?

All bases below quality score 35 were trimmed? That is a very stringent criterion. You would lose a lot of data if not a single base below 35 quality score is retained post-trimming. Why did the authors choose to use this?

TZ - please expand the abbreviation.

bwa aln gives aligned files, not counts. What was used to generate the raw counts? Also, bwa is not a splicing aware aligner. Authors should use Bowtie instead to do this, which may prove to be a better alternative.

All the figures need to be improved. In all figures, the sequence in question should be highlighted / boxed to make it easier for the reader to follow what is being talked about.

Figure 5 legend - 'shows the repeat'. which repeat? Authors should clearly mention that there are 2

contiguous repeats of the same 39aa sequence in the transcript C55368_G1_I3. Also, there are 2 different reasons mentioned why the assembler could not merge the 2 transcripts in question. Please clarify which is the case.

GitHub repository
The README file requires substantial work. Some of the commands are quite confusing, comments are not clear, and perl scripts are not to be found. The numerous manual steps make the tool virtually un-useable.

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Competing Interests: No competing interests were disclosed.

Referee Report 28 January 2016

doi:10.5256/f1000research.7788.r12066

Michael Love
Department of Biostatistics, Harvard TH Chan School of Public Health, Boston, MA, USA

I do not have expertise in transcript assembly, but I can comment on the general readability and usability of the article and tool suite.

1. As with the report from Dr. Charoensawan, I was expecting that the tool suite would be more integrated and documented than the collection of Perl scripts on the Github page. Nevertheless, the set of examples shown in the article I think are useful at showing the kinds of errors which arise from transcript assembly and presumably it is easy to use the scripts to identify such examples. The README is currently very minimal for a tool suite / integrated workflow, looking more like a set of comments above code rather than proper documentation of a tool suite. I would recommend changing from README to README.md, using Github markdown e.g. enclosing the commands in backticks, re-writing the comments as full sentences/paragraphs, separating the different steps by sub-headings, etc. A little effort here will make the landing page much more appealing. Also the documentation on the Github page should provide detailed information on the expected inputs and outputs.

2. The colors in Figure 1 make the text a bit hard to read. As Figure 1 is often where many readers will go to understand what you are doing, you would benefit from making the colors lighter so the text is easier to read, and removing unnecessary shading, 3D effects, etc.

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Competing Interests: No competing interests were disclosed.

Author Response 30 Jan 2016

Sandeep Chakraborty, Tata Institute of Fundamental Research, India

We thank you for taking the time to review this paper, and for your comments.

As we have mentioned previously in response to Dr Charoensawan, this manuscript is not meant to be a software article. The primary reason for this is that the flow is not push-button. One simple goal of this paper was to highlight downstream checks that can correct and improve the results of a heuristic assembler like Trinity (which is bound to have certain limitations).

An enhanced version of the merging algorithm (which struck us later on) is to check whether the E-value of the merged transcript decreases when BLAST'ed as compared to the two transcripts being merged. This definitely would point to a non-merged assembly.

Also, most of the methods described here are reasonably simple to code.
In time, as we figure out how to automate the scripts better, we will certainly incorporate your suggestions.

We will revise the manuscript with a simpler and less distracting version of Figure 1.

Competing Interests: No competing interests were disclosed.No competing interests were disclosed.

Referee Report 04 January 2016

doi:10.5256/f1000research.7788.r11300

Varodom Charoensawan
Mahidol University, Bangkok, Thailand

The authors have addressed most of my previous comments.

However, I still have one reservation on the use of *Arabidopsis* "proteome" (instead of of publicly available transcriptomes) as a benchmark for walnut transcripts found, in the section "Identifying highly transcribed genes that are not present in the *Arabidopsis* proteome". It would be useful if the authors could clarify this.

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Competing Interests: No competing interests were disclosed.

Author Response 05 Jan 2016

Sandeep Chakraborty, Tata Institute of Fundamental Research, India

Dear Dr Charoensawan,
We would like to thank you once again for critically reviewing, and accepting the revised version.

As for the Table 2, which mentions the "Identifying highly transcribed genes that are not present in the Arabidopsis proteom
e" that you have found inadequately explained, we would like to specify that

1) This set of transcripts were first chosen as they have high expression levels (Table 1).
This table already shows the homology of the ORF's of these transcripts to the BLAST 'nr' database (apart from the first transcript, all other have some degree of homologous counterparts).
2) Next, Arabidopsis was chosen on purpose since (as we mention in the text) "it fails to accumulate wood".
Our intention was to extricate differences in the transcripts which probably define the wood quality of walnut. Choosing other proteomes that included other wood generating plants would not suffice to find such transcripts.

We will try to rephrase this part to make it more lucid when we make another version (it would be too small a change for a new version, otherwise).

best wishes,
Sandeep

Competing Interests: No competing interests were disclosed.

Version 1

Referee Report 05 October 2015

doi:10.5256/f1000research.7105.r10335

Varodom Charoensawan
Mahidol University, Bangkok, Thailand

Chakraborty and coworkers proposed a new platform for analysing transcriptomic data from RNA-seq (YeATS -Yet Another Tool Suite for analyzing RNA-seq derived transcriptome). The key feature of the tool highlighted by the authors is error estimation and correction of assembled transcripts, which is performed by analysing ORFs predicted in each transcript and merging of transcripts. This error-filtering step is supposedly missing in most other existing tools to date. In addition, YeATS is able to perform other common RNA-seq analytic tasks, such as transcript abundance estimation.

From the point of view of a frequent user of NGS tools, rather than a developer, I can see that such a tool can be useful for improving transcript assembly and estimation, especially in organisms with no or poorly annotated genomes. However, there are a number of points that, to me, would improve the tool and the article, and it would be great if the authors could address/clarify. I would be happy to discuss this further if any of my comments are not clear.

- It would be nice to include a performance evaluation of this new platform against existing tools, or with vs' without the transcript error correction step by YeATS. One way to do this might be to take an existing RNA-seq dataset from a well-annotated organism such as *Arabidopsis* as a gold standard, and perform transcript assembly-estimation with and without correction by YeATS, and compare this to the transcript estimation using genomic information (e.g. by mapping reads to annotated transcriptomes/genomes). Does YeATS indeed improve the coverage and specificity of transcript estimation, for instance?

- Along the same lines as the comment above, it would be useful if the authors could comment on the time and/or computing resources required to perform the correction step. Also, are the accuracy and computing resources dependent on the read lengths and/or sequencing platforms? (Or is it intended for Illumina reads as used in the example?).

- Both the source code of YeATS and the data set used to illustrate its usage have been deposited and described at the end of the article. However, to this reviewer's understanding, there is a set of Perl scripts deposited to Github, but it is still not clear to me how the tool/workflow should be implemented. The README does not seem to describe this. Could the author point out if there is already a guideline or documentation on how to use or integrate YeATS into an existing NGS workflow, if that already exists?

- To my understanding, the input of YeATS is a set of assembled transcripts performed by other tools (e.g. Trinity). However, this step was not clearly described in the "*in vitro* methods" section on Page 5. Instead, it seems the trimmed reads were directly aligned to *J. regia* transcriptome, which is somewhat confusing. Could you please clarify these?

- The authors described the genes as highly "transcribed" in walnut (according to RNA-seq from this study?) that are not present in Arabiodopsis "proteome". I found these to be slightly disconnected.

Minor comments:

- Figure 1: Is the "no" label between the boxes "Choose longest ORF" to "Gene annotation" necessary?
- Page 3, 2nd column, line 12: modify the text to "often in distinct regions of the transcript.." for clarity?
- Page 5, 1st column: There were ~24K "of" such transcripts
- Figure 6's legend: These ORF"s"

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Competing Interests: No competing interests were disclosed.

Author Response 13 Oct 2015

Sandeep Chakraborty, Tata Institute of Fundamental Research, India

We would like to thank you for taking the time to review this paper. Please find our responses below.

Chakraborty and coworkers proposed a new platform for analysing transcriptomic data from RNA-seq (YeATS -Yet Another Tool Suite for analyzing RNA-seq derived transcriptome). The key feature of the tool highlighted by the authors is error estimation and correction of assembled transcripts, which is performed by analysing ORFs predicted in each transcript and merging of transcripts. This error-filtering step is supposedly missing in most other existing tools to date. In addition, YeATS is able to perform other common RNA-seq analytic tasks, such as transcript abundance estimation. From the point

of view of a frequent user of NGS tools, rather than a developer, I can see that such a tool can be useful for improving transcript assembly and estimation, especially in organisms with no or poorly annotated genomes.

We appreciate your positive comments, and the possibility of value addition by the suggested methodology for existing NGS flows. We believe that this is the first attempt to associate the key information encoded by transcripts within ORFs to assess the accuracy of the assembly.

However, there are a number of points that, to me, would improve the tool and the article, and it would be great if the authors could address/clarify. I would be happy to discuss this further if any of my comments are not clear. It would be nice to include a performance evaluation of this new platform against existing tools, or with vs without the transcript error correction step by YeATS. One way to do this might be to take an existing RNA-seq dataset from a well-annotated organism such as Arabidopsis as a gold standard, and perform transcript assembly-estimation with and without correction by YeATS, and compare this to the transcript estimation using genomic information (e.g. by mapping reads to annotated transcriptomes/genomes). Does YeATS indeed improve the coverage and specificity of transcript estimation, for instance?

YeATS evaluates the accuracy of a transcriptome, but it is dependent on downstream tools (like MAKER) to use this for proper annotation of the genes. Thus, there are no existing tools that we could compare it with directly. A highly curated database like the Arabidopsis would not be a fair comparison, since it might have been annotated looking at several data points. However, we have extensively used the YeATS pipeline in processing the newly sequenced walnut genome (manuscript in review), and established erroneous assembly for several genes of interest. The transcriptome from several other tissues were included in the genome study. Interestingly, the 5% error estimate remained the same.

Along the same lines as the comment above, it would be useful if the authors could comment on the time and/or computing resources required to perform the correction step. Also, are the accuracy and computing resources dependent on the read lengths and/or sequencing platforms? (Or is it intended for Illumina reads as used in the example?).

The run times for most of the processing required in YeATS is a few hours on a 16 GB, 16-core machine, barring the search for homologies in the BLAST 'nr' database, which can be time-intensive for a comprehensive search. This search can be significantly accelerated when the organism under investigation has well-annotated protein databases (as in the current case), much in lines of the newly introduced SMARTBLAST (http://blast.ncbi.nlm.nih.gov/smartblast/), to run times under a day. Run times are dependent on the number of transcripts only, since the input to YeATS is an assembled transcriptome from a tool like Trinity. We have included this information in the manuscript.

Both the source code of YeATS and the data set used to illustrate its usage have been deposited and described at the end of the article. However, to this reviewers understanding, there is a set of Perl scripts deposited to Github, but it is still not clear to me how the tool/workflow should be implemented. The README does not seem to describe this. Could the author point out if there is already a guideline or documentation on how to use or integrate YeATS into an existing NGS workflow, if that already exists?

We have provided a README that describes the step in the YeATS workflow. However, this is not a push-button methodology, and goes through several steps, each of which is dependent on the previous step. Also, we have used custom schedulers, and thus several steps need to be adjusted depending on available resources. For example, the number of parallel jobs and the time-step between each submission is controlled through a custom-script. Thus, we have provided the key algorithms in detail in the paper for any developer to easily replicate our results. Furthermore, we are enhancing several of the programs based on more sophisticated algorithms (like using kmers, compression of data, etc). A proper release of this software will require some more time, but this manuscript was not meant to be a software article.

To my understanding, the input of YeATS is a set of assembled transcripts performed by other tools (e.g. Trinity). However, this step was not clearly described in the in vitro methods section on Page 5. Instead, it seems the trimmed reads were directly aligned to J. regia transcriptome, which is somewhat confusing. Could you please clarify these?

The input of YeATS is indeed a set of assembled transcripts performed by other tools like Trinity. We have modified the methods section to clarify this.

The authors described the genes as highly transcribed in walnut (according to RNA-seq from this study?) that are not present in Arabiodopsis proteome. I found these to be slightly disconnected.

We agree that these results are slightly disconnected to the general narrative of this paper, which focuses on post-assembly methodologies to assess the accuracy of assembled transcripts. However, these are interesting results that emerge during the analysis of the transcriptome of the transition zone of walnut, which has been obtained for the first time. And thus, though this may be of interest to researchers in the field, there is too little data to spin-off another paper to publish these findings.

Minor comments:
Figure 1: Is the no label between the boxes Choose longest ORF to Gene annotation necessary?

We have changed the label 'no' to 'unannotated'. Long ORFs that do not have have significant matches are probably uncharacterized genes, and the genome could be annotated accordingly (although the annotation of novel genes is another problem not addressed in the current paper).

Page 3, 2nd column, line 12: modify the text to often in distinct regions of the transcript.. for clarity?

We have clarified this: 'The ORFs map to different fragments of the same protein. This points to an error in the sequencing or the assembly, which breaks down the contiguous ORF into two fragments.'

Page 5, 1st column: There were 24K of such transcripts Figure 6s legend: These ORFs

We have made these modifications. Once again, we are thankful for your insightful comments, and hope to have addressed your concerns.

Competing Interests: No competing interests were disclosed.